

DOCUMENT RESUME

ED 284 875

TM 870 348

AUTHOR Forster, Fred; Karr, Chad
TITLE The Accuracy of the Fixed Parameter Method of Item Bank Calibration.
PUB DATE [87]
NOTE 13p.
PUB TYPE Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Item Analysis; *Item Banks; Junior High Schools; Language Tests; Latent Trait Theory; Mathematical Models; Mathematics Tests; Reading Tests; *Sample Size; Standardized Tests; Test Construction; Test Items; *Test Validity
IDENTIFIERS *Calibration; *Fixed Parameter Approach; Rasch Model; Tests of Individual Performance

ABSTRACT

This study investigates two fixed parameter models for calibrating test items for an item bank under local school district control. Mathematics, reading, and language tests administered to students (Grades 7-9) provided the criterion information. Model I used calibrations from the Northwest Evaluation Association Item Bank to calibrate new items on the criterion-referenced test. Six random samples of 150, 200, 250, and 300 students were drawn for the condition of "fixing" 10, 20, or 30 items of a 40-item test. There was no discernible effect of fixing increasing numbers of test items. Model II used student scores from the Rasch scaled Tests of Individual Performance (TIP) for its item calibrations. Calibration accuracy was compared for samples of 100, 150, or 200 students drawn without replacement. A sample size of 200 was recommended for accurate item calibration. Model II applied well, even when the calibrating test and the test to be calibrated were given at different times of the year, with appropriate adjustment for average achievement differences in the time interval. Model II is recommended as an improved item analysis technique because: (1) it uses a mathematically independent criterion; and (2) it can be related to any Rasch scale. (LPG)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

Fred Forster

THE ACCURACY OF THE FIXED PARAMETER
METHOD OF ITEM BANK CALIBRATION

Fred Forster
Microsystems for Education & Business, Inc.
109 N. Main Ave., Gresham, OR 97030

Chad Karr
Department of Psychology
Portland State University
P.O. Box 751, Portland, OR 97207

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

Introduction

The Northwest Evaluation Association (NWEA) item banks are a landmark development for local school districts, because they put the critical elements of curriculum aligned standardized test development under district control. With their 10-year history of analysis and re-analysis, the NWEA item banks are probably the most thoroughly researched testing resource available for assessing academic achievement in the elementary schools. As a result, several districts in the Northwest routinely develop tests to blueprint specifications, engineer desired measurement characteristics into their tests, and predict the appropriateness and performance of their tests even before they are administered to students. These districts have shown that the development of quality curriculum aligned testing is within the grasp of all local districts without dependence on test publishers or centralized agencies.

The research and practical engineering necessary to use the Rasch model to develop large scale calibrated item banks was pioneered by Portland (Oregon) Public Schools. This development included the creation of linking network design, linking analysis, and linking cross validation analysis to ensure the integrity of the resulting scales. This work is discussed in Ingebo, Arneklev, and Gee (1976), Ingebo (1976), and Karr (1976).

The fixed parameter model was developed to increase the efficiency of procedures for building calibrated item banks, without sacrificing accuracy or cross validation. The unique concept embodied in the fixed parameter approach is the use of test information from one situation to analyze the results from a second, independent testing situation. For example, the fixed parameter makes it possible to calibrate and item analyze each question on a new test using previously established student scores on a standardized test or calibrations of items drawn from an item bank. (For a discussion of the application of the fixed parameter model see Forster, 1982a; Forster, 1982b; and Forster & Karr, 1986.)

This study was designed to determine the accuracy of the fixed parameter model in calculating the calibrations for new items, to determine the time limitations involved in the application of the model, and to explore the use of the fixed model in item analysis.

The Fixed Parameter Model

The name for the fixed parameter model rises from the use of previous information to "fix" one or more of the parameters estimated in a latent trait analysis. For example, the traditional Rasch procedure uses current test information to estimate both the achievement level corresponding to each possible raw score and the calibration

ED284875

TM 870 348

corresponding to the p-value for each item. It then proceeds to refine these estimates to establish the best match to current test data. Using the fixed parameter procedure, the achievement estimates can be established from item bank calibrations or previous standardized achievement scores which are independent of the current data. This capability makes it possible to calculate the calibration of each item directly from the previous information. In this way the fixed parameter model provides a mathematical basis for isolating and estimating latent trait parameters, one at a time.

This study used two different fixed parameter models. As illustrated in Figure 1, the Fixed Parameter Model I uses previously determined item calibrations to calculate student achievement levels based on the portion of the test drawn from the item bank. These achievement levels are then used to calculate the calibration of each new item directly. As illustrated in Figure 2, the Fixed Parameter Model II uses Rasch unit achievement scores for students from a standardized test to calculate the calibration for each new item directly. This model is mathematically more powerful than Model I since each item is calibrated on independent information drawn from a separate test administration. Errors or weaknesses in any of the items on the new test do not affect the calibration of the other items on that test.

Research Questions

This study addresses the following questions:

1. What is the smallest acceptable sample size needed to calibrate new items using a calibrating test (Fixed Parameter Model II)?
2. What is the smallest number of previously calibrated items which can be used to calibrate new items (Fixed Parameter Model I)?
3. What is the effect of the time difference between the administration of the calibrating test and the test to be calibrated (Fixed Parameter Model II)?
4. How can item analysis be extended using a calibrating test (Fixed Parameter Model II)?

Methodology

The data used in this study were drawn from tests taken by students in grades 7 to 9 in the Hillsboro Public Schools. These data sources included:

1. Mid-year and end-of-semester tests in mathematics administered in January 1985. These tests included 7th grade general mathematics, 7th grade pre-algebra, 8th grade general mathematics, Algebra I, and geometry.
2. End-of-year semester tests in mathematics administered in June 1985. These tests included the same subjects as listed above.
3. The Tests of Individual Performance (TIP) standardized Rasch calibrated tests in reading, language, and mathematics administered in April 1985.



Figure 1

FIXED PARAMETER MODEL I



Figure 2

FIXED PARAMETER MODEL II

4. The TIP scores from September 1985.

The TIP scores were matched to the end-of-semester test data and served as the independent "calibrating" test to make it possible to apply the Fixed Parameter Model II.

The test data were scanned and scored by a microcomputer system using error checking procedures which required the operator to examine and check every answer sheet on which one or more multiple marks were detected or more than 10% of the questions were left unanswered. Since this methodology is more conservative than usual scanning/scoring procedures, it increases the accuracy of the data and, consequently, the precision of data analyses.

The data were analyzed by the fixed parameter procedures based on the unconditional maximum likelihood equations developed by Wright (Wright & Panchapakesan, 1969 and Wright & Stone, 1979) and adapted by Forster (1978).

Results

The results of the four parts of the study are summarized below in relation to each research question addressed.

Question 1: What is the smallest acceptable sample size needed to calibrate new items using a calibrating test (Fixed Parameter Model II)?

This question was addressed using the results from the spring TIP testing as the calibrating test for the spring end-of-semester tests. The calibrations were calculated using populations ranging from 419 to 650 students. Random samples of 100 students, 150 students, and 200 students were drawn without replacement until the population was exhausted, and then the procedure was repeated until eight samples had been drawn. (This procedure provided "relative independence" among the samples used in the study.) The calibrations for the samples were compared to those for the population, and both the average difference and difference in absolute value were calculated. There were eight samples for each group size for each of four tests. The differences were summed across items and across tests.

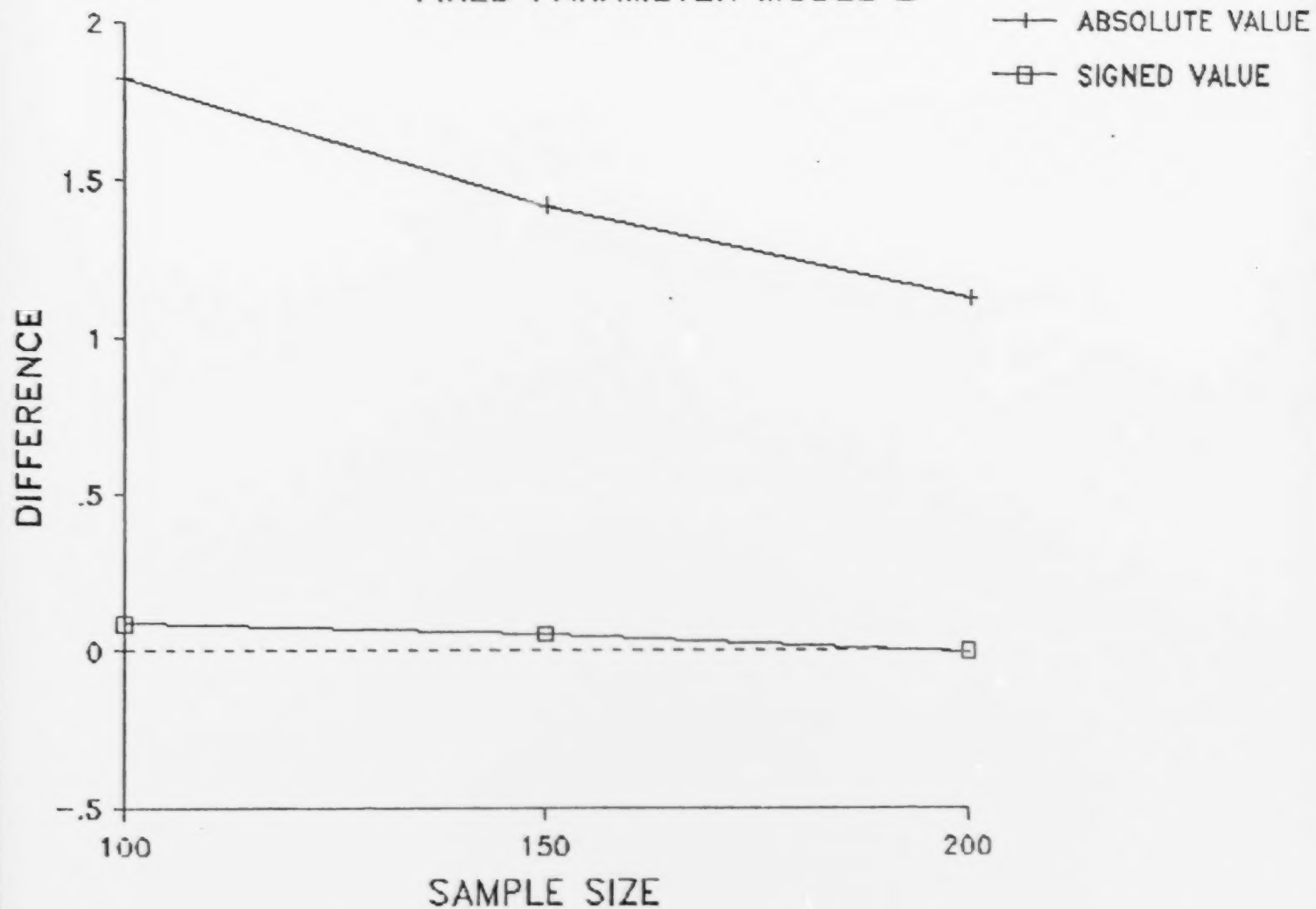
As shown in Figure 3, the average differences were close to zero, indicating that the discrepancies in calibration were symmetrically distributed and that the scaling for the samples closely approximated the total population. The average absolute value differences were smaller than previously established for the Rasch procedure, ranging from 1.83 for the samples of size 100 to 1.12 for the samples of size 200. Based on these data, samples of 200 or even 150 students appear sufficient to provide accurate item calibrations using the Fixed Parameter Model II.

Question 2: What is the smallest number of previously calibrated items which can be used to calibrate new items (Fixed Parameter Model I)?

It was not possible to attack this question in full generality since item bank calibrations were not available for the tests being analyzed. Instead, the study "fixed" the calibrations for items based on the Model II calibration for the total population. For this reason, the differences reported here should be considered a lower bound for the results expected when actually using an item bank.

Figure 3

AVERAGE DIFFERENCE IN CALIBRATION FIXED PARAMETER MODEL 2



This question was addressed using the results from the 7th grade mathematics spring end-of-semester test. The calibrations were calculated for the full sample of 467 students. A set number of these calibrations were established as "fixed" based on the full sample calibrations, and the remaining items were calibrated for each sampling. Six random samples of 150, 200, 250, and 300 students were drawn for the condition of "fixing" 10 items, 20 items, and 30 items. (There were 40 items in the test.) The calibrations for the samples were compared to those for the population, and both the average difference and difference in absolute value were calculated. The differences were summed across items and averaged.

As shown in Figure 4, the average differences were small but not symmetric. While there was a trend for the differences to be smaller for larger sample sizes, there was no predictable trend for the effect of "fixing" increasing numbers of the items in the test.

As shown in Figure 5, the absolute differences were small and decreased for the larger sample sizes. Surprisingly, there was no discernible effect of "fixing" increasing numbers of items in the test.

Question 3: What is the effect of the time difference between the administration of the calibrating test and the test to be calibrated (Fixed Parameter Model II)?

This question was addressed using the results from both of the spring end-of-semester tests. These tests were calibrated using the Fixed Parameter Model II and the TIP scores for students in spring 1985 and fall 1985.

As shown in Table 1, the shift in calibration using the fall 1985 calibrating tests was approximately equal to the difference in the average TIP achievement between the two periods. These data indicate that the calibrating test and the test to be calibrated can be administered at different times of the year by making a simple arithmetic adjustment equal to the difference in average RIT achievement levels to adjust the resulting calibrations.

Question 4: How can item analysis be extended using a calibrating test (Fixed Parameter Model II)?

This question was addressed using the results from the fall 1985 TIP testing and a criterion referenced test in reading administered in fall 1985. The Fixed Parameter Model II was used to analyze the test using reading and then language TIP scores. As shown in Table 2, there were significant differences among the point biserials calculated using the CRT raw score, the TIP reading score, and the TIP language score. The TIP point biserials were lower than the raw score point biserials since the effect of intercorrelation between the criterion and item data had been removed. Most interesting, some of the TIP point biserials differed significantly as a result of the alignment of the item to the reading and language curriculum scales.

Discussion

The results from this study establish a bench mark of 200 students or more as an optimum sample size for the use of the Fixed Parameter Model II.

Figure 4

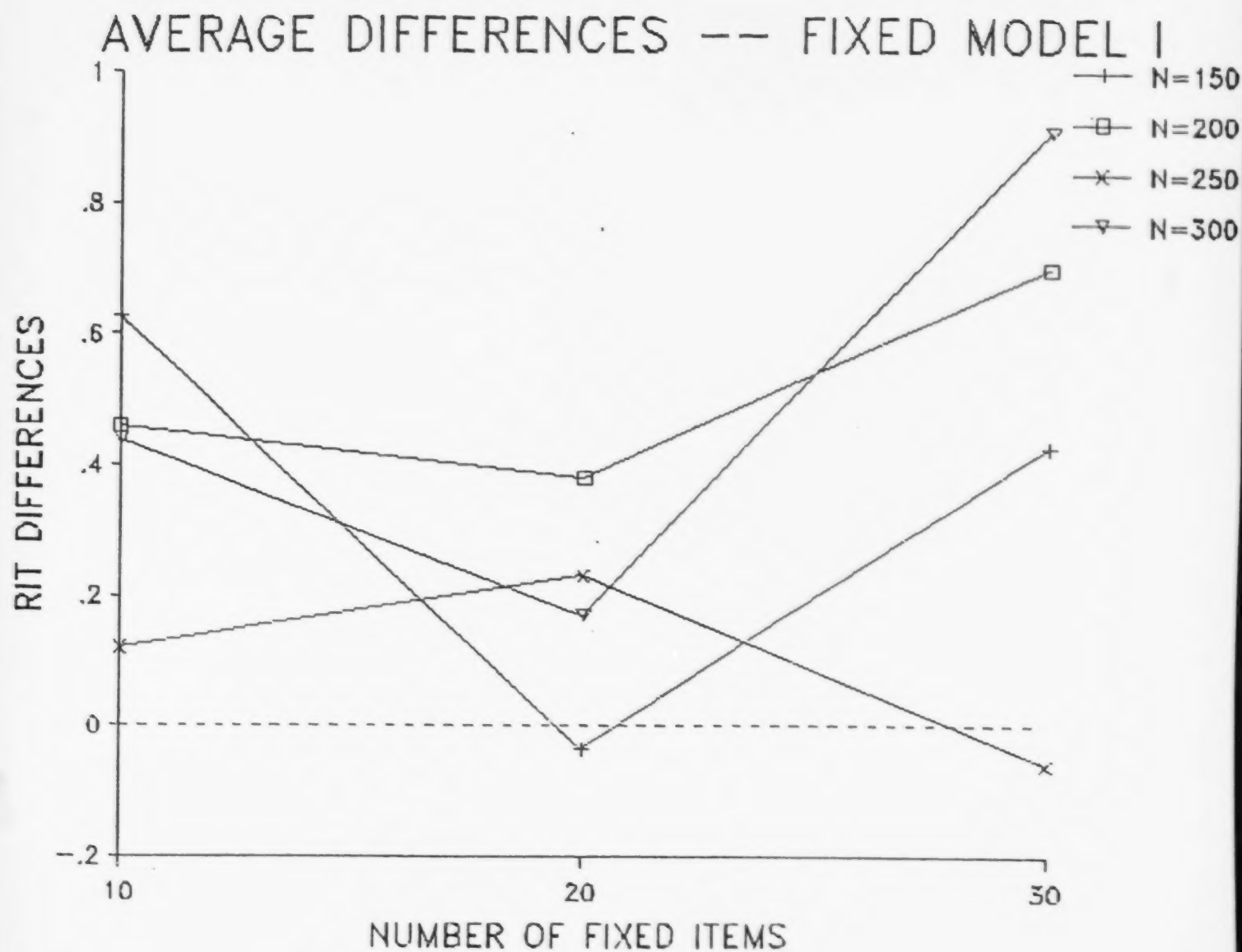


Figure 5

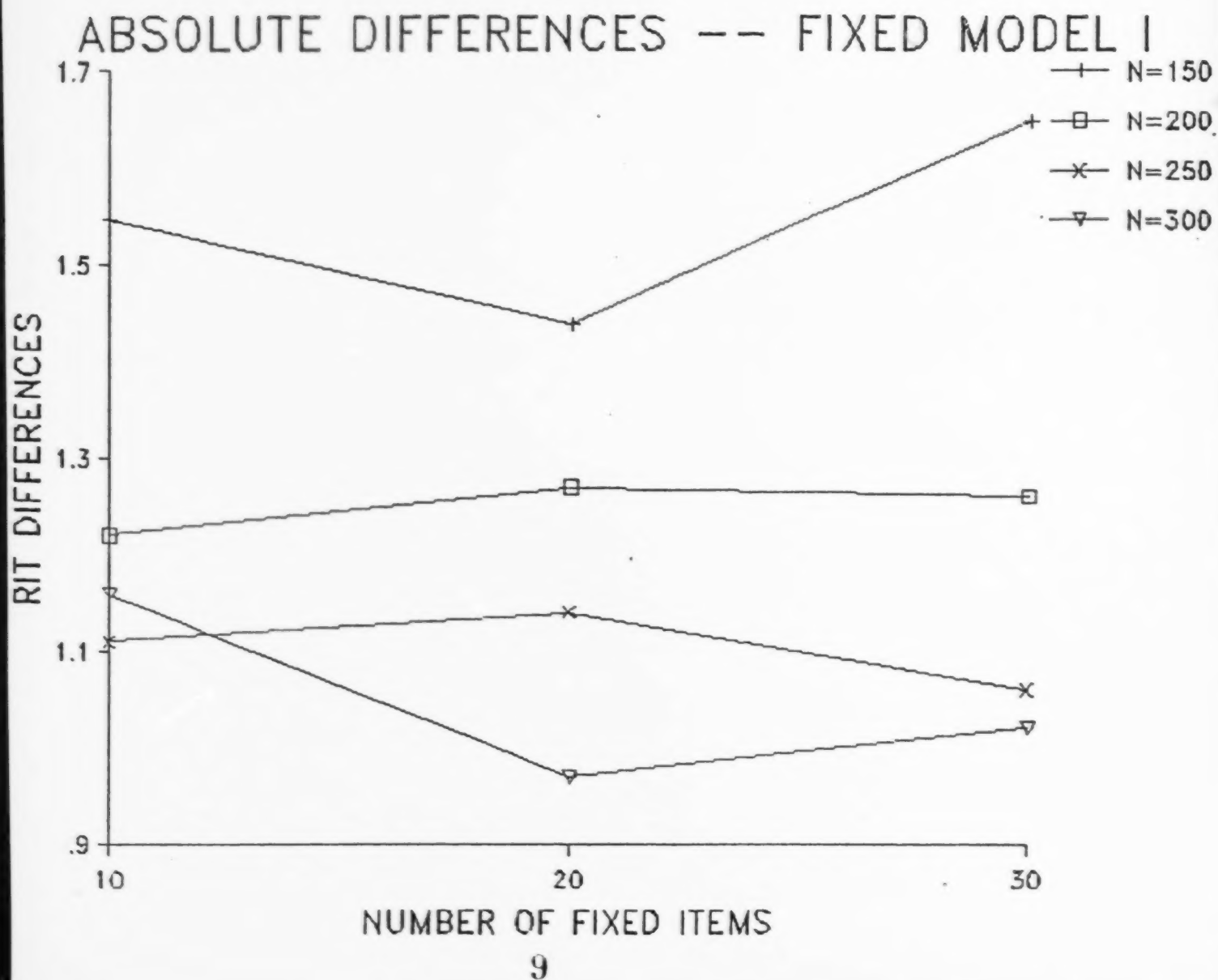


Table 1
The Effect of Time on the Fixed Parameter Model II

	SPRING 85		FALL 85		DIFFERENCE	
	AVERAGE TIP ACH	AVERAGE CRT CAL	AVERAGE TIP ACH	AVERAGE CRT CAL	AVERAGE TIP ACH	AVERAGE CRT CAL
TEST 1	221.8	221.4	223.7	223.9	1.9	2.5
TEST 2	226.6	219.9	227.4	220.7	0.8	0.8
TEST 3	237.5	234.9	239.3	237.1	1.8	2.2
TEST 4	246.9	244.1	246.3	242.8	-0.6	-1.3

Table 2
Comparison of Point Biserials

ITEM	CRT RAW SCORE	READING TIP SCORE	LANGUAGE TIP SCORE
1	0.35	0.16	0.21
2	0.20	0.20	0.10*
3	0.36	0.19	0.25
4	0.45	0.32	0.28
5	0.39	0.27	0.22
6	0.31	0.15	0.13
7	0.22	0.12	0.09
8	0.24	0.11	0.15
9	0.37	0.29	0.24
10	0.36	0.29	0.27
11	0.39	0.23	0.30
12	0.44	0.35	0.30
13	0.26	0.19	0.19
14	0.37	0.18	0.20
15	0.12	0.03	0.08
16	0.37	0.27	0.21
17	0.52	0.42	0.39
18	0.46	0.40	0.31*
19	0.33	0.30	0.29
20	0.42	0.39	0.35
21	0.27	0.14	0.16
22	0.16	0.03	0.11*
23	0.32	0.20	0.22
24	0.18	0.12	0.08
25	0.37	0.25	0.18
26	0.48	0.36	0.28
27	0.36	0.30	0.29
28	0.22	0.10	0.11
29	0.30	0.19	0.11*
30	0.38	0.31	0.31
31	0.29	0.25	0.19
32	0.38	0.32	0.26
33	0.10	-0.01	-0.09*
34	0.01	-0.04	-0.02
35	0.25	0.16	0.21
36	0.16	0.07	0.13
37	0.33	0.22	0.22
38	0.11	0.04	0.07
39	0.13	0.13	0.14
40	0.15	0.07	0.05

* Difference in TIP point biserials is statistically significant (Fisher's Z transformation)

The results from this study indicate that sample size is related to the accuracy of the Fixed Parameter Model I. Surprisingly, the accuracy does not appear to have a simple relationship to the number of items fixed. We expected a direct relationship between the number of items fixed and the accuracy of calibration. The next step in this research will be to determine the effect of fixing different items and then to identify the differential effects. This research is also incomplete since it was not possible to fix items based on the use of item bank values. Future research is needed to relate calibrations drawn from item banks to the stability of the calibration of new items using the Fixed Parameter Model I.

The results of this study indicate the Fixed Parameter Model II can be applied even when the calibrating test and the test to be calibrated are given at different times of the year, if calibrations are adjusted to reflect the difference in the average RIT achievement of the students between the two points in time.

Finally, the results of this study indicate that the Fixed Parameter Model II promises to be a valuable item analysis technique. Since it uses a mathematically independent criterion, it removes the spurious correlation between criterion variable and item information present in current item analysis techniques. Since it can be related to any Rasch scale, it provides information about the fit of items between school subjects.

References

Forster, F. (1978). Mathematics of the Rasch model. Northwest Evaluation Association Rasch Monograph Series, 23-37.

Forster, F. (1982a). The fixed parameter model: a more accurate and flexible tool for item calibration. Occasional Paper No. 23, Portland Public Schools, Portland, Oregon.

Forster, F. (1982b). The fixed parameter model. Paper presented at the annual meeting of the Canadian Educational Research Association Conference, Nova Scotia.

Forster, F., & Karr, C. (1986). Build your own calibrated item bank. Paper presented at the Fourteenth Annual Pacific Northwest Research and Evaluation Conference, Vancouver, B.C., Canada.

Ingebo, G. S. (1976). How to link tests to form an item pool. Occasional Paper No. 6, Portland Public Schools, Portland, Oregon.

Ingebo, G. S., Arneklev, B., & Gee, D. (1976). Optimum range of difficulty for linking items. Northwest Evaluation Association Journal, 2(1), 23-32.

Karr, C. (1976). Evaluation by triangulation. Occasional Paper No. 4, Portland Public Schools, Portland, Oregon.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago: MESA Press.